The EMBO Journal Vol.2 No.8 pp.1357 – 1365, 1983

Evidence for a repeating cross-β sheet structure in the adenovirus fibre

N.M. Green*, N.G. Wrigley, W.C. Russell, S.R. Martin and A.D. McLachlan[1]

National Institute for Medical Research, Mill Hill, London NW7 1AA, and [1]Medical Research Council Laboratory for Molecular Biology, Hills Road, Cambridge, UK

Communicated by A.E. Smith
Received on 2 May 1983; revised on 1 June 1983

The amino acid sequence of the adenovirus fibre protein reveals an approximately repeating motif of 15 residues. A diagonal comparison matrix established that these repeats extended from residue 43 to residue 400 of the 581 residue sequence. Assignment of secondary structure combined with model building showed that each 15-residue segment contained two short β-strands and two β-bends, one of which incorporated an extra residue in a β-bulge of the Gx type. The 44 strands together gave a long (210 Å) narrow, amphipathic β-sheet, which could be stabilised by dimer formation to give the shaft of the fibre. The knob could arise from a dimer of the C-terminal 180 residue segment, predicted to be an 8 – 10 stranded β-sandwich. This model is consistent with the electron micrographs of the fibre and it was supported by measurements of c.d. and of electron diffraction from microcrystals. The latter gave a pair of wide angle arcs, corresponding to a repeat of 4.7 Å, oriented appropriately for a cross-β structure. The relation of this structure to globular structures is discussed and a folding pathway is proposed. In its general features the structure resembles that proposed for the tail fibre of bacteriophage T4.

Key words: adenovirus fibre/protein structure prediction/electron diffraction/cross-β structure/repetitive sequence

Introduction

The icosahedral capsid of adenovirus consists of 252 capsomeres of which 12 are situated at the 5-fold vertices and have therefore been designated pentons (Valentine and Pereira, 1965). The pentons are composed of a penton base anchored in the capsid and a non-covalently attached fibre protein terminating in a knob. The fibre appears to be responsible for attachment to cells and for the haemagglutinating activity of the virus (Ginsberg, 1979). The length of the fibre is characteristic of a given subtype, and varies from ~10 nm for human adenovirus type 3 and to almost 50 nm for some fowl adenoviruses (Norrby, 1969; Laver *et al.*, 1971; Gelderblom and Maichle-Lauppe, 1982). In the latter case, most of the serotypes elaborate a pair of fibres, one short and one long, emanating from each penton base.

SDS-polyacrylamide gel electrophoresis has shown that the fibre polypeptide has a mol. wt. of 60 000 – 64 000 and this appears to be true even for serotypes of differing fibre lengths (Wadell, 1979), although in most analyses the assignment of a polypeptide species to fibre has been based only on analogy with the well characterised serotypes of human adenoviruses type 2 and 5. The native fibre protein has been reported to have a mol. wt. of ~200 000 (Dorsett and Ginsberg, 1975; Sundquist *et al.*, 1973) suggesting a trimeric structure, but much lower values of 150 000 – 160 000 have been reported recently (Devaux *et al.*, 1982), consistent with a dimeric structure. The fibre is coded by a single gene and sequencing the adenovirus type 2 fibre gene region has revealed a single open reading frame which codes for a polypeptide of mol. wt. 61 925 (Hérissé and Galibert, 1981; Hérissé *et al.*, 1981).

We have analysed the amino acid sequence derived from the above studies and have discerned a 15-residue segment which repeats approximately, over >300 residues. This forms the basis for a description of the overall morphology of the fibre in terms of the amino acid sequence. New experimental evidence is presented in support of the proposed structure. This evidence was obtained with fibre prepared from adenovirus type 5, since it gave crystals suitable for electron diffraction. We can justify the use of these results to support a structure based on the sequence of adenovirus type 2 since both types are closely related members of subgroup C of human adenoviruses (Green *et al.*, 1979) and there is no significant difference in the length of the fibres (Norrby, 1969).

Results

Statistical analysis of the amino acid sequence

The secondary structure predicted by the method of Chou and Fasman (1978) suggested that the main central segment of the protein consisted of short β-strands alternating with β-bends. It was not at first possible to evaluate the prediction because of the many overlapping bends and the shortness of the strands (4 – 6 residues). However, a regular periodicity of the hydrophobic residues was noted between 90 and 160 with maxima and minima approximately every eight residues as well as the regular appearance of a proline at 15-residue intervals. The maxima corresponded to sequences such as Leu-Ser-Val, often predicted as β-strands, and the minima were usually predicted to be β-bends. We do not show details of the predicted structure, but the general conclusions are illustrated in Figure 3.

It seemed likely that this region would form a long, narrow, antiparallel β-sheet with a hydrophobic face and a polar face but before attempting to deduce the secondary structure of the whole sequence we analysed it for repeating segments.

To clarify our proposal we use a structural notation for the 15 positions of the basic repeating sequence (Figure 1). The unit is believed to contain two short β-strands, A and B, linked in anti-parallel fashion by two bends. The B strand is followed by a normal β-bend of four residues (b_1, b_2, b_3, b_4) while the A strand is followed by a five-residue bend (a_1, a_2, a_3, a_4, a_5) which can be regarded as a normal β-bend with the addition of an extra residue (often proline) in position a_5. Positions A_1, A_3, B_1 and B_3 are almost always hydrophobic and a_2 is often hydrophobic. We made the arbitrary decision to commence each repeat at residue b_3, in the middle of the 'b' bend, since there are some irregularities which are most simply described as insertions at this point.

We examined the amino acid sequence systematically by the comparison matrix method (McLachlan, 1971, 1972) using a scoring system based on observed mutation rates in

*To whom reprint requests should be sent.

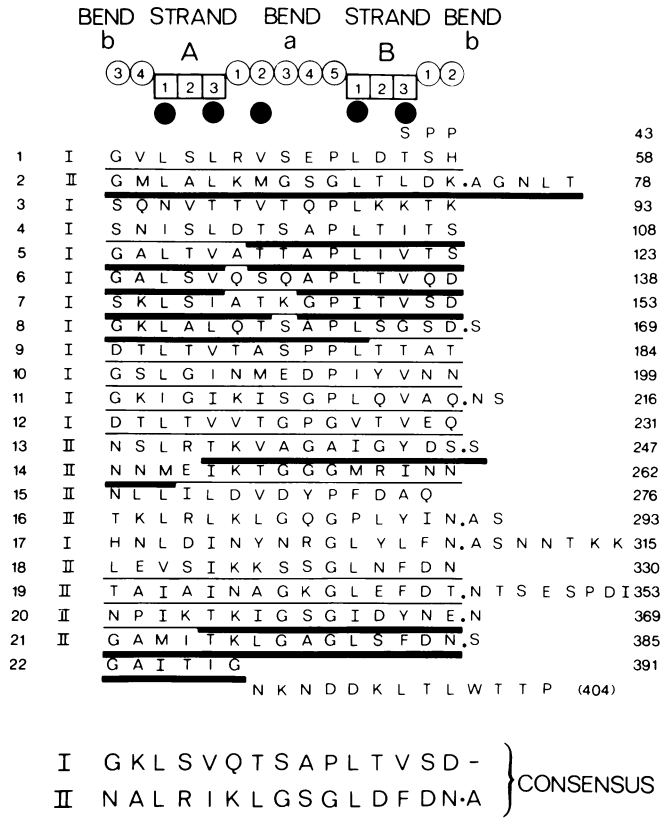

Fig. 1. Repeating segments from the shaft region of the fibre protein. The structural pattern of two strands and two bends in each segment is shown above, with black circles designating hydrophobic sites. The residue number of the last amino acid in each segment is shown on the right side (assuming that the N-terminal formyl methionine has been cleaved off). The segment number and group type (I or II) are on the left. The clear repeats are underlined and heavy underlinings mark the most closely related sections of amino acids (Table II). Note the clusters of polar residues inserted in some of the bends. Below are shown typical consensus sequences for the two families of repeats.

homologous proteins (Dayhoff, 1978). The significance of a high score between two segments was tested against the calculated 'double matching probability' (McLachlan, 1971). This is the probability that a score at least as high as the observed one will be found in the comparison between two randomly chosen segments from a large sample with the same average composition (see Table I).

The matrix in Figure 2 illustrates the main features of the pattern, which includes 21 or 22 recognisable 15-residue repeat segments from positions 44–400 of the sequence. The central block of repeats, residues 94–216, is particularly strong and regular and contains the segments numbered (4–11). There is also a block of four more strong repeats (18–21) at the end of the regular region. These latter repeats are seen to be homologous to segments (11–14) and so all the repeats are interconnected. Another prominent feature of the pattern, especially in the central block, is a weaker 'false repeat' along diagonals midway between the principal diagonals. This is a consequence of the alternations of polar and non-polar side chains in the strands A and B, which leads to a high score when the A strand of one segment is aligned with a B strand seven or eight residues further on in another segment.

The assignment of the segments in Figure 1 was made by placing the strongest repeats first and assembling the weaker fragments into a consistent pattern. The repeats are strong

Table I. Distribution of comparison scores in the repeating region of the fibre sequence

Score (S)	Observed number S or above	Expected frequency	Observed frequency	Ratio observed/ expected
Amino acid comparisions (span 15)				
163	2784	1.1×10^{-2}	2.8×10^{-2}	2.6
170	1022	1.2×10^{-3}	1.1×10^{-2}	9.2
176	453	1.4×10^{-4}	4.7×10^{-3}	34.6
182	164	1.3×10^{-5}	1.7×10^{-3}	130.8
187	50	1.4×10^{-6}	5.2×10^{-4}	371.4
192	7	1.3×10^{-7}	7.2×10^{-5}	553.8
195	3	2.9×10^{-8}	3.1×10^{-5}	1069.0
DNA comparisons (span 21 codons)				
25	3415	1.7×10^{-2}	3.5×10^{-2}	2.1
28	730	2.1×10^{-3}	7.5×10^{-3}	3.6
31	154	1.5×10^{-4}	1.6×10^{-3}	10.7
33	31	2.0×10^{-5}	3.2×10^{-4}	16.0
35	8	2.1×10^{-6}	8.3×10^{-5}	39.5
36	2	6.3×10^{-7}	2.1×10^{-5}	33.3

Scores were taken from the complete set of 96 580 independent span comparisons between all pairs of sequence fragments in residues 1–440 of the amino acid sequence or codons 1–440 of the nucleic acids sequence. Amino acid scoring system based on Dayhoff (1978). Nucleic acids scored as 0–3 identities per codon. All codons compared in the actual reading frame.

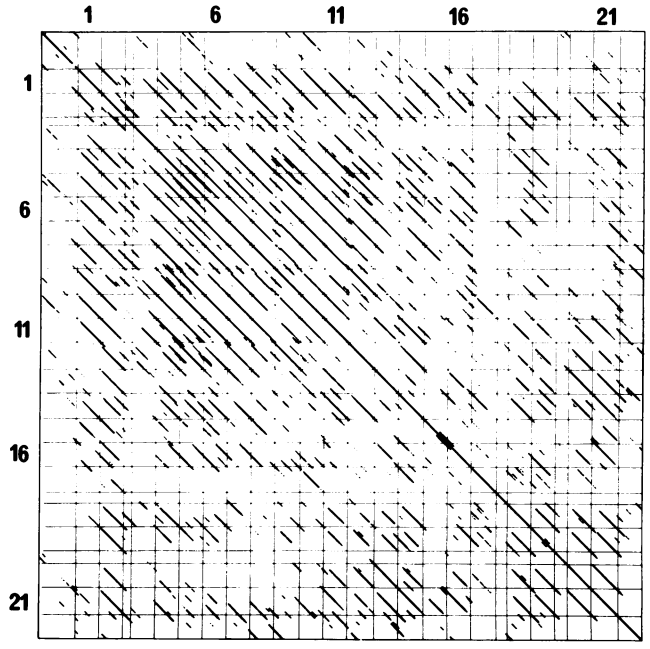

Fig. 2. Sequence comparison matrix for the amino acid residues 21–440. Each point on the diagonals marks the mid-point of two 15 residue sequence fragments (measured across the columns and down the rows) whose comparison score is above the 1% double matching probability level. The thickened lines have scores above the 0.5% level. Grid lines are drawn at the ends of the repeat segments defined in Figure 1. The diagram is best read by scanning across a row corresponding to a particular segment and observing the frequency of strong matches to other segments. The lengths of the diagonals give the coherence length of each alignment.

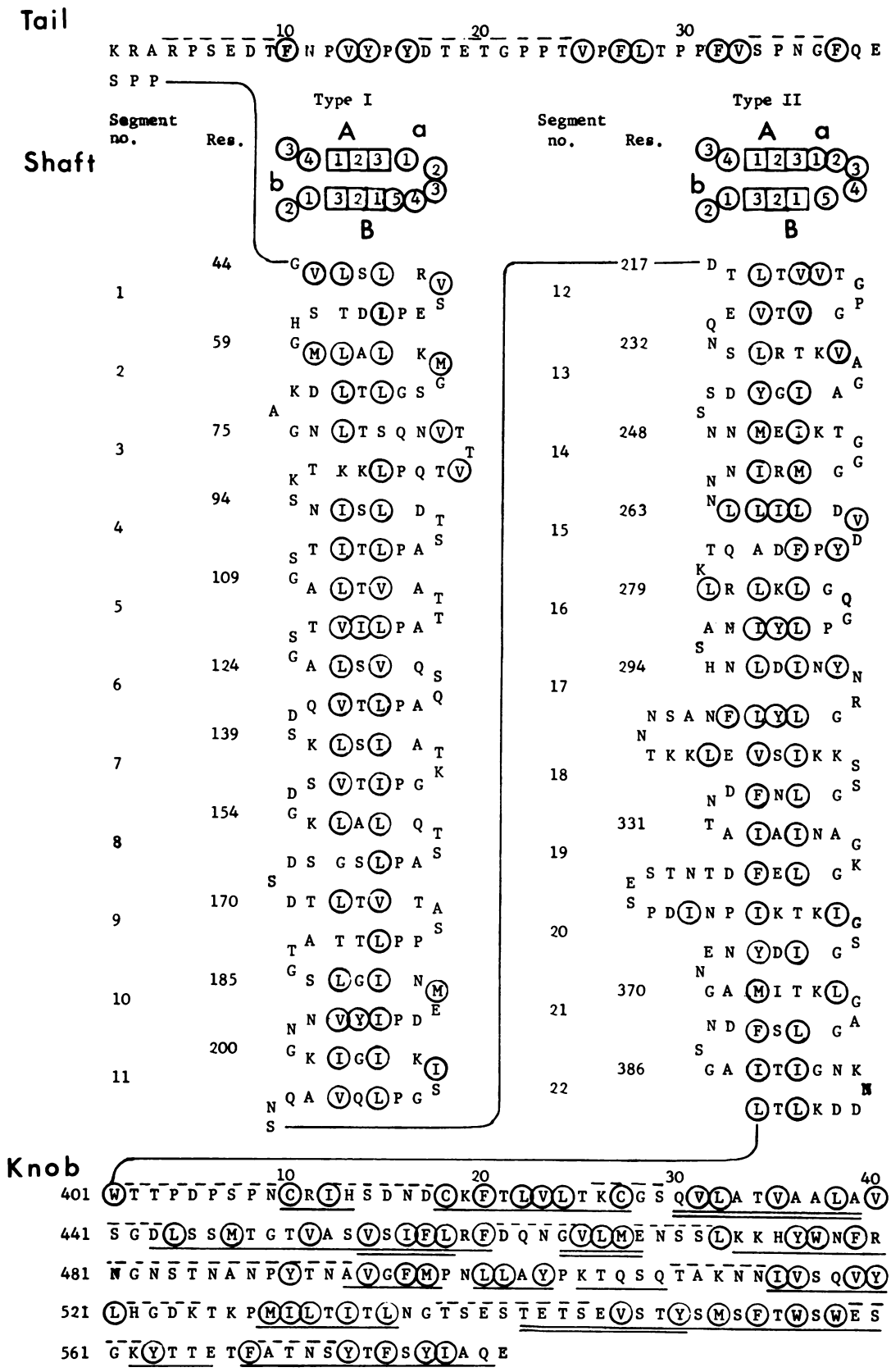

Fig. 3. Structural organisation of the subunit of the fibre protein. Each subunit comprises a short N-terminal tail separated from a distal knob by 350 residues of repetitive sequence which forms the shaft. The predicted secondary structure of the tail and head are indicated as β-strand , β-bend or α-helix. In the shaft region all but six of the bends were predicted as β-bends and ~70% of the β-strands were correctly predicted. All large hydrophobic residues are ringed and most of those in the shaft face upwards. Note that 90% of the residues in positions A_1, A_3, B_1, B_3 are hydrophobic in contrast to only 8% of the 132 residues which alternate with them. *All segments but 2 in the first column are of type I and have a proline at a_5, probably forming residue 2 of a Gx bulge (see Figure 6). **All segments but 12 and 17 in the second column are of type II and many have a glycine position a_5 which may form either residue 2 or (as illustrated) residue x of a Gx bulge.

Table II. Closely related regions of the amino acid sequence

Segment numbers	Position of first region	Position of second region
Amino acid comparisons (span 15, matching probability $<10^{-6}$)		
2,21	Gly 59 - Lys 73	Gly 370 - Asn 384
4,5	Thr 100 - Val 113	Thr 115 - Val 128
6,7	Ala 132 - Thr 145	Gly 147 - Thr 160
4,7	Thr 106 - Leu 120	Thr 150 - Leu 164
13,20	Thr 236 - Met 250	Thr 358 - Met 372
20,21	Thr 358 - Met 372	Thr 374 - Ile 388

and unambiguous in 16 of the segments, numbered 1, 2, 4–12, 14 and 18–21. Segment 13 is identified by its close relationship with 2, 20 and 21, while segment 16 is consistently related to 2, 13 and 21. Segments 17 and 10 are closely related, but the two weakest segments, 3 and 15, are tentative. The last fragmentary segment, 22, is terminated by a proline-rich sequence Pro-Asp-Pro-Ser-Pro-Asn following residue 403. The final pattern requires some insertions at the ends of regular segments. The asignment of segment 16 is not unique, as an almost equally good match is obtained by starting the A strand at Leu 281 and shortening the proline bend by one residue. This second choice is used in the structural diagram (Figure 3).

The observed degree of similarity between amino acid segments is far too strong to have arisen by chance and so the repeats have probably evolved by a series of gene duplications. Table I shows that the numbers of high scores recorded are many times greater than expected. The trend is particularly marked among the higher scores. For example, the highest score expected for a random sequence (of length 440) is 182. It should occur once, but actually occurs 164 times. The chance of obtaining the highest observed score of 195 even once is only 10^{-3}, but it occurs three times. Table II lists the most closely related segments, with comparison scores of 187 or more. Notice that the segments in Table I overlap one another, so that the number of independent segments listed in Table II is smaller.

Statistical analysis of the DNA sequence

We also analysed the repeats directly from the gene, comparing all possible ranges of codons in their correct reading frame. The codon comparison matrix was scored by counting identical bases. The DNA repeat pattern is far less significant and much harder to interpret than the amino acid pattern and we found it necessary to use a longer span of 21 codons (63 bases). The strongest DNA repeats then largely confirm the amino acid ones. The weakness of the DNA repeats suggests that the DNA sequence itself, in this gene, is only subject to an indirect evolutionary constraint, the need to code for a fibre with appropriate structure.

Evolution of the repeats

We attempted to construct an evolutionary history for the 21 repeated segments (Fitch, 1977). Since the repeats form a cyclic pattern the analysis is difficult. There is no natural genetic division point between segments, and it is quite possible that the proposed gene duplications occurred across different boundaries in different segments. The relationships in Table II suggest that the most closely related segments fall into two groups. Group I segments, related to segments 4, 5, 6, 7, 8 typically have proline in position a_5 and Ile or Val in position B_3. Group II segments, related to 2, 18, 19, 20, 21 typically have Gly or Ala at a_5 and a_3, Lys at a_1 Tyr or Phe at B_3 and an enlarged b bend. Segments were assigned to the two classes according to scores from an amino acid family comparison matrix (McLachlan, 1972), in which segments 4–8 were used as standards for group I and segments 2 and 18–21 for Group II. Figure 1 shows that the two groups are clearly segregated within the sequence and that segments 11–14 form an intermediate set with similarities to both I and II.

These observations suggest that evolution of the repeats has largely proceeded in a linear fashion along the sequence, as if an ancestral group I repeat unit based on segments 4 and 5 has gradually multiplied itself and changed, first into the intermediate family 11–14 and then into the nucleus of a new group II. Equally, the sequence of events could have been reversed, since there is no way to order the relationships in time. We have not been able to proceed further and build up an evolutionary tree as there are too many ambiguities.

Prediction of secondary structure

The regularity of the repeats combined with the uniform appearance of the shaft of the fibre in electron micrographs suggested that these features correspond and that the 15-residue segments have a common secondary structure. The regularity enables one to reject misleading 'noise' in the Chou and Fasman prediction and to propose the more firmly based structure shown in Figure 3. The limits of the shaft region were established by our analysis of the repeats. Within the shaft, all but six of the 44 bends were predicted as such and so were most of the short strands. Helices, however, were predicted in segments 1, 2, 7, 8, 15, 16 and 17, but, in view of the uniformity of the shaft these segments probably form β-strands. The 'b' bends were typical β-bends of four residues in which b_1 and b_4 continued the hydrogen bonding pattern of the β-strands. The extra residue in the five-residue bend at position a_5 was most commonly a proline situated between a typical β-bend sequence and the beginning of the B strand.

The assembly of the strands and bends into a long, narrow β-sheet is shown in Figure 3 together with the predicted secondary structure of the tail (1–43) and knob (400–581) regions. The N terminus was rich in proline (25%) and contained several β-bends. Its configuration may be determined by its interaction with the penton base. The knob appears to consist predominantly of β-strands, much longer than those of the shaft, joined by β-bends and α-helical segments. It is likely that this region of the molecule forms an anti-parallel β-sandwich of 8–10 strands (c.f., Richardson, 1981). Most of the strands show the alternation of polar and hydrophobic residues characteristic of β-sandwiches. A few, which contain runs of hydrophobic residues, might be involved in inter-subunit contacts.

Relation of the predicted structure to the macroscopic appearance of the fibre

The long, narrow β-sheet structure which we propose for the shaft has a hydrophobic face and a polar face. It would not be stable in isolation, but a dimer with a 2-fold axis coincident with the fibre axis would be a plausible, stable structure. The 44 β-strands would give the shaft a length of 210 Å (44 x 4.7 Å, Geddes *et al.*, 1968). The knob would then form a globular dimer of mol. wt. 40 000, corresponding to a diameter of 45 Å. The N-terminal tail might add

Table III. Amino acid composition of adenovirus fibre

	Type 5 Mol%[a]		Mol/mol[d]	Type 2 Mol/mol[e]
Asp	13.4	(12.2)	78	76
Thr	11.5[b]	(11.4)	67	68
Ser	8.8[b]	(8.1)	51	62
Glu	7.1	(6.3)	41	36
Pro	4.4	(6.2)	36	34
Gly	8.5	(8.9)	49	46
Ala	7.5	(8.1)	49	37
Val	5.4	(5.9)	31	33
Met	1.2	(1.3)	7	11
Ile	3.6	(4.5)	21	32
Leu	12.1	(13.3)	70	56
Tyr	2.6	(2.3)	15	17
Phe	3.1	(3.8)	18	17
His	1.2	(1.2)	7	5
Lys	5.1	(6.1)	30	34
Arg	1.5	(1.4)	9	10
Cys	N.D.	(0.4)	(2)	3
Trp	1.1[c]	(1.0)	6	4

[a]Mean of four determinations. Figures in parentheses from Dorsett and Ginsburg (1975).
[b]Uncorrected for destruction during hydrolysis.
[c]Calculated from the u.v. absorption (Melamed and Green, 1963).
[d]Assuming 581 residues total.
[e]Hérissé *et al.* (1981), assuming removal of N-terminal FMet.

another 10–20 Å to the length to give a total of 270 Å. This is close to the observed dimensions of type 2 and type 5 fibres (Valentine and Pereira, 1965; Norrby, 1969;. and Figure 5 below).

We built duplicate models of repeats 8–12 from both Nicholson models and CPK space-filling units. The two halves could be fitted closely together with non-polar faces in contact to give a rigid compact structure of cross section 18 Å x 23 Å with almost all the large hydrophobic residues buried in the interior of the sandwich.

Experimental results

The proposed structure is consistent with the overall dimensions of the fibre. We consider next some experiments designed to check the following further conclusions, which can be drawn from the model. (i) The c.d. spectrum should be characteristic of protein with ~50% of β-sheet and very little α-helix. (ii) Microcrystals in which the fibre molecules are oriented parallel to the crystal axis (Mautner and Pereira, 1971) should give wide angle electron diffraction with reflections at 4.7 Å in line with the crystal axis and characteristic of a cross-β structure (Earnshaw *et al.*, 1979). (iii) The fibre should be dimeric. (iv) The N-terminal end should interact with the penton base.

C.d.

All the experimental results were obtained with fibre protein from adenovirus type 5. Its amino acid composition is compared with the results of Dorsett and Ginsburg (1975) and with the composition of the type 2 fibre in Table III. The agreement between the two sets of analyses of the type 5 fibre was reasonably good and showed that the type 5 protein had

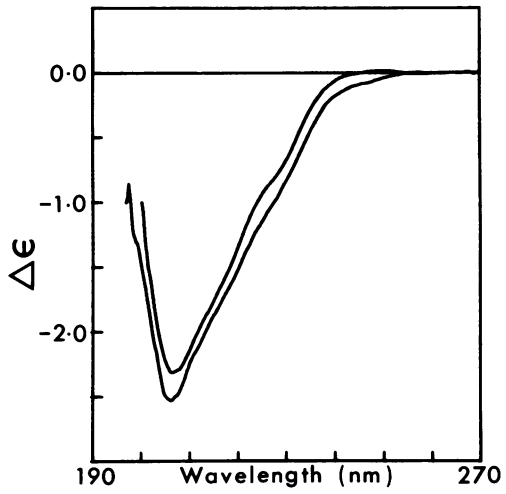

Fig. 4. C.d. spectrum of adenovirus fibre. The two spectra were obtained from two different preparations of the protein. Each is an average of six scans.

Table IV. Comparison of secondary structure from the c.d. with that assigned from the model.

	α-helix	β-sheet	β-bend	Random
Preparation 1 c.d.	16.2	37.6	23.3	22.7
Preparation 2 c.d.	19.7	30.9	22.1	27.3
Calculated from Figure 3[a]	5	55	21	21

The c.d. spectrum was analysed by the method of Provencher and Glockner (1981).
[a]In the shaft region ten residues from each segment were assigned to β-structure. Similarly, throughout the whole structure, two residues of each β-bend were assigned to β-bend configuration and two to β-sheet.

a significantly higher alanine and leucine and a lower isoleucine content than that from type 2. The results are consistent with a considerable degree of homology between the two proteins and our results are also close to those obtained by Boulanger and Loucheux (1972) with adenovirus type 2.

The c.d. spectra of two different preparations of the fibre protein (Figure 4) show a trough at 208 nm and no significant positive dichroism above 195 nm. Analysis of the spectrum in terms of contributing secondary structures (Provencher and Glöckner, 1981) showed a predominance of β-sheet as expected, but also suggested 15–20% (~100 residues) of helical contribution (Table IV). This is several times higher than implied by the structure proposed in Figure 3. Although the fit of the calculated curve to the experimental curve was good, it is possible that the set of proteins used for matching the c.d. spectrum may not have been appropriate for a protein of such unusual structure. The work of Woody (1974) has shown that certain kinds of β-bend can give negative dichroism at 208 nm which would lead to an overestimation of α-helix. The presence of a large number of bends, many containing an unusually located proline residue, might in part have been responsible for the unexpectedly high negative dichroism at 208 nm.

Electron diffraction by crystals of the fibre

The fibre protein has been shown to form microcrystals (Mautner and Pereira, 1971). Two forms were obtained and

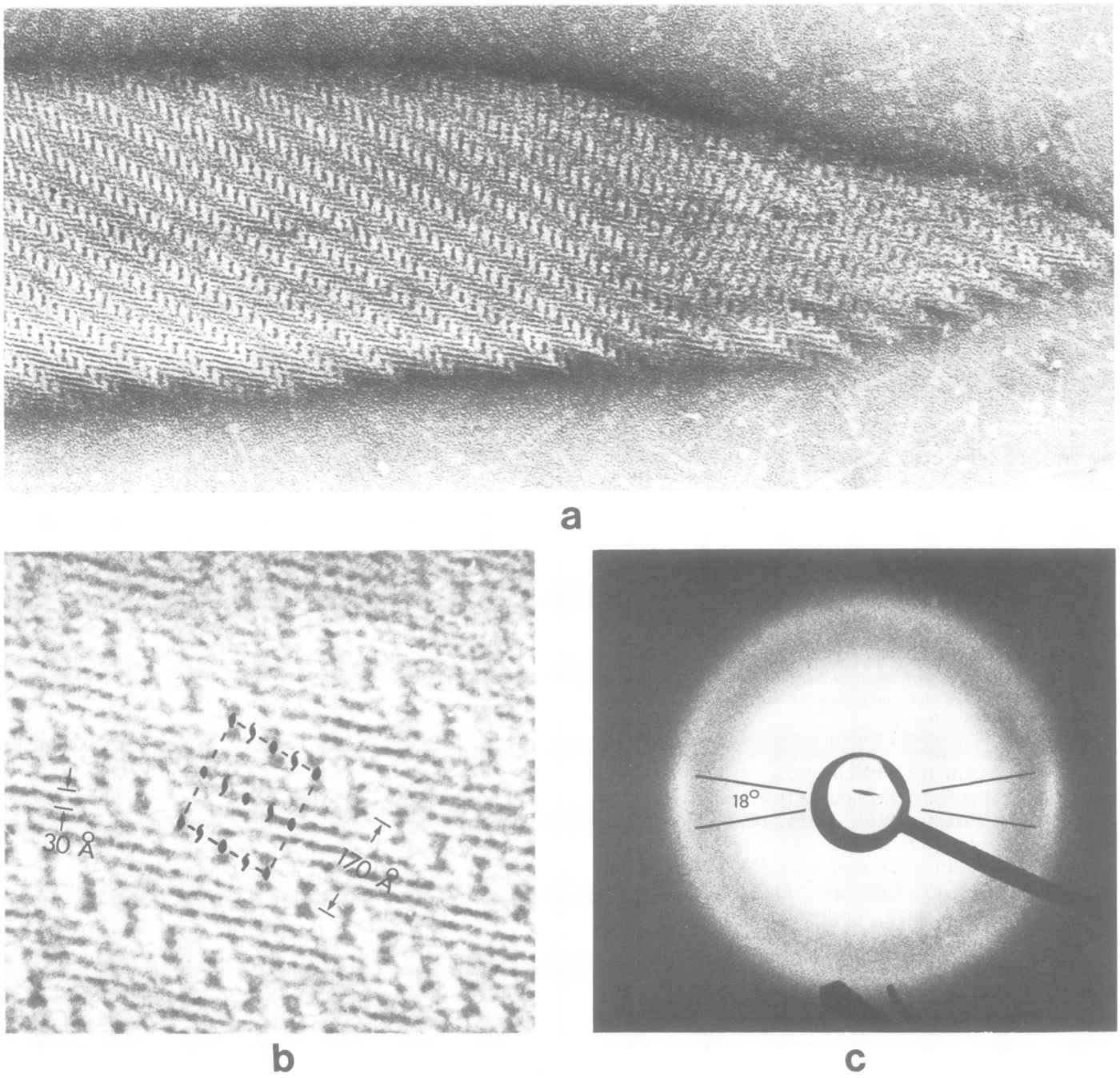

Fig. 5. Electron microscopy of crystals of the adenovirus fibre. **(a)** Crystal negatively stained with sodium silicotungstate. Note single molecules of the fibre in the background (x 300 000). **(b)** The high magnification image shows oblique rows of approximate ellipses linked by ladders of thin fibres. The pattern can be analysed in terms of anti-parallel pairs of fibre molecules. The members of each pair are superimposed and related by a vertical 2-fold screw axis near the middle of the shaft. A second screw axis between the knobs of two anti-parallel pairs generates the ellipses. Combination of these screw axes with ordinary 2-fold axes, as shown, generates the whole crystal with monoclinic space group C2, having its unique axis perpendicular to the micrograph. The unit cell was chosen to give the closest approach to an orthogonal section. This interpretation, made with the assistance of Professor T. Blundell, is supported by electron micrographs showing ends of crystals where the alternation of knobs and tails of fibre molecules is clearly visible. **(c)** Electron diffraction from a crystal similar to **(a)**. Unstained. The pair of arcs, corresponding to a spacing of 4.7 Å, subtend an angle of ~18° and lie almost on the crystal axis (small image at centre).

one of these (Figure 5a) was thin enough to give a wide angle diffraction pattern (Figure 5b), which however, was very short lived and was never seen on the screen. The reflections consisted of two short arcs subtending an angle of 18 ± 2°. The relative azimuth of these arcs to the long axis of the crystal was 5 ± 2°, corresponding roughly to the azimuth of the fibre to the crystal axis. This angular relation was observed in all of the 30 crystals for which we recorded diffraction patterns. In each case the arcs corresponded to a spacing of 4.7 Å, calibrated against the gold crystal spacing, d_{111} = 2.35 Å. This provides proof that much of the polypeptide chain of the fibre protein forms a β-sheet with the chains running perpendicular to the fibre axis, as suggested by our model.

The dimensions of 100 fibre molecules were measured on negatively stained electron micrographs of free fibres. The

mean length was 308 ± 27 Å and the mean diameter of the knob was 54 ± 10 Å, in good agreement with earlier results (Valentine and Pereira, 1965; Norrby, 1969).

The subunits of the fibre protein

A thorough study of the penton and its constituents (Devaux *et al.*, 1982) gave a mol. wt. of 156 000, half way between the expected figures for dimer and trimer. Recently, Devaux *et al.* (see Acknowledgements) have produced strong evidence for the dimeric structure by determining the number of monomers in the unit cell of crystals of adenovirus type 2.

Polarity of the fibre

Although the model agrees well with the observed dimensions, its polarity is the reverse of that recently suggested by Boudin and Boulanger (1982) on the basis of the effect of digestion of the fibre with carboxypeptidase Y. Their conclusion is open to doubt, since the released amino acids were not identified and carboxypeptidase Y often contains traces of endopeptidases, which could remove material from the N terminus during the long (8 h) incubation period.

Discussion

There are several features of the shaft of the adenovirus fibre protein which put the proposed structure on a firmer basis than that of most predicted structures. The repetitive sequence and the unusual geometry of the shaft impose narrow limits to the configuration of the peptide chain, which are further narrowed by the crystallographic and spectroscopic evidence for a β-sheet structure. The sequences in the alternating bend and strand segments are close to ideal for the proposed structure and the models which we built were compact and had no implausible stereochemical features. Nevertheless, the structure does show two features which have not been commonly observed in established protein structures and these are worth further consideration. The features are (i) the five-residue bends, often containing a proline residue at a_5, and (ii) the β-sandwich structure, in which the hydrophobic interior is unusual in being intermolecular rather than intramolecular.

The five-residue proline bends

The simplest way of forming the cross β-sheet of the type proposed here would be to link the strands with four-residue turns only, as envisaged by Geddes *et al.* (1968) in their analysis of the cross-β structure of Chrysopa silk. However, this is possible only when the strands contain an even number of residues (four in Chrysopa silk). The geometry of a β-strand is such that combination of strands containing an odd number of residues with β-bends leads to cyclic structures. For example, three-residue strands would give a cyclic tetradecapeptide and single residue 'strands' would give a cyclic decapeptide (e.g., gramicidin S).

The five-residue bend provides an escape from this structural difficulty as well as giving desirable stabilising features. This can be seen by reference to Figure 6, where the hypothetical addition of a cyclic tetradecapeptide to a hairpin loop structure is illustrated. Strands A and B, hydrophobic residues uppermost, are shown forming a 14-residue cyclic tetradecapeptide. This is added to the hairpin by breaking the bond between residues B_1 and a_4. The hydrogen-bonded ring, containing residues a_2, a_3 and a_4 is then rotated about the C_α-CO and C_α-NH bonds of residue a_1. The carbonyl group of a_4 can be connected to the NH of the proline (a_5), inserted at the N-terminus of the hairpin without any strain, giving a β-bend which is tilted upwards at an angle to the β-sheet. The first side chain of the turn (a_1) remains on the polar face of the sheet but the succeeding three side chains (a_2, a_3 and a_4) are rotated onto the hydrophobic face. The plausibility of this model is confirmed by inspection of the sequence (Figure 1) from which it can be seen that while a_1 is usually polar (85%), only 40% of the residues in positions a_2 and a_4 are polar. In contrast, >75% of the corresponding residues of the 'b' bends are polar.

When the proposed dimeric sandwich is formed, the tilted proline bends seal the edges of the sheets and their hydro-

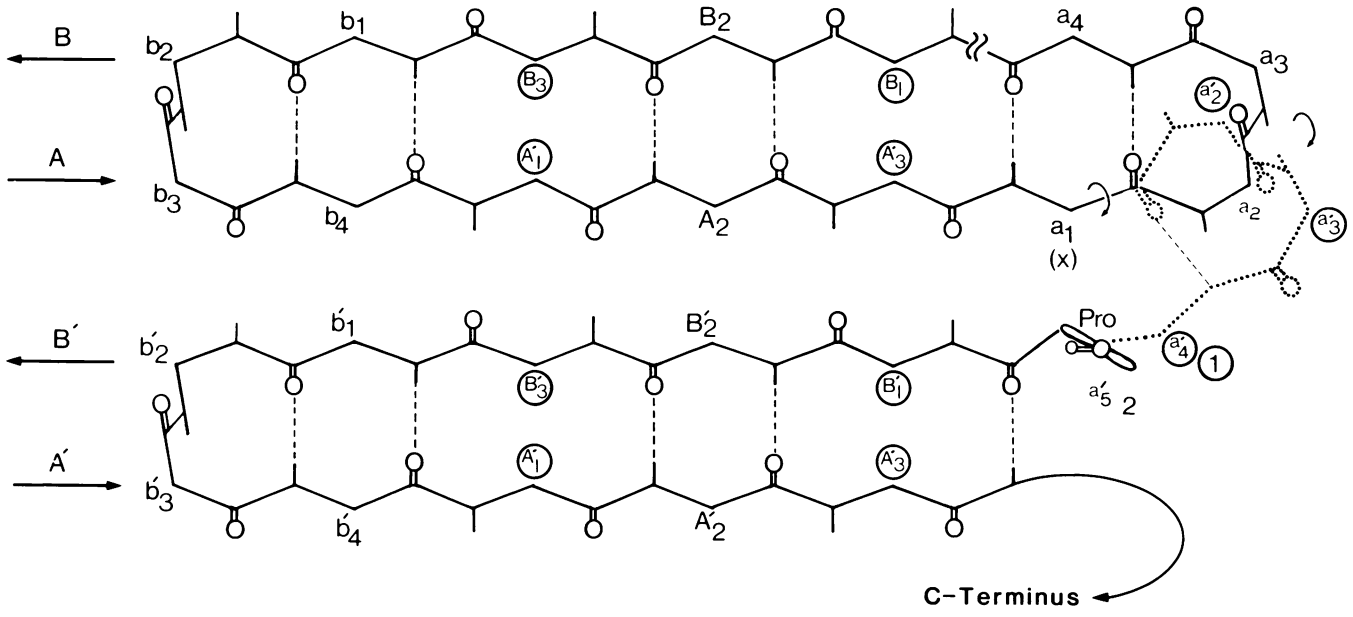

Fig. 6. Addition of a hairpin loop to a cleaved cyclic tetradecapeptide. Only the encircled residues lie above the plane of the paper. After cleavage between B_1 and a_4 the residues a_2, a_3 and a_4 of the β-bend are rotated together about the two bonds of a_1, until the carboxyl group of a_4 can link to the proline a_5, inserted before the B′ strand of the hairpin loop. The β-bulge present in the resulting five residue bend is designated by X, 1 and 2 (Richardson *et al.*, 1978).

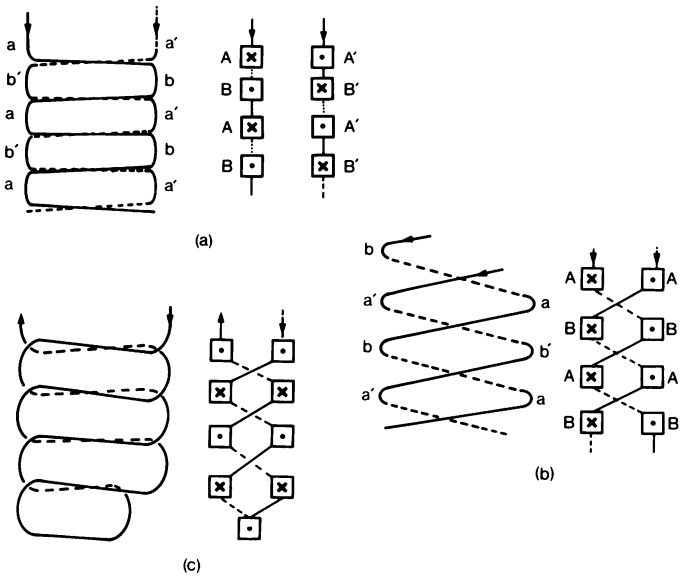

Fig. 7. Comparison of the strand connections in two hypothetical adenovirus fibre structures with those found in tomato bushy stunt virus. (a) Side by side anti-parallel β-sheet. (b) Parallel double helix. (c) Idealised Swiss roll topology as found in tomato bushy stunt virus (Richardson, 1981). The view of the structure in the topological diagram on the right of each Figure is rotated at 90° about a vertical axis in the plane of the paper relative to the main figure. X represents the N-terminus and • the C-terminus of each strand. Since the strands in both (b) and (c) are wound into right-handed double helices, there should be a natural right handed crossover between strands on opposite faces of the sandwich. This is the opposite of the usual crossover in globular proteins (7c and text). The 'crossover' referred to in this context is not the same as the 'crossover' of connections between successive strands in a β-sheet. In globular proteins these two types of crossover have opposing handedness (Richardson, 1981).

phobic residues provide extra internal bulk, giving a compact structure with a hydrophobic core and a polar exterior. The only slightly unfavourable feature of the structure is the burial of the proline carbonyl group without any complementary hydrogen bond donor.

An alternative description of the bend in terms of β-bulges relates this structure to known features of globular proteins. The effect of inserting an extra residue next to position 1 or 4 of a β-bend is to give a structure resembling one of the 'narrow' β-bulges described by Richardson *et al.* (1978), since residues a_1 and a_4 have approximately a normal β-configuration. Because the insertion is accompanied by a 180° rotation of the bend (Figure 6), the bulge will be of the Gx type, in which residues a_1, a_4 and a_5 correspond to residues x, 1 and 2 of the Richardson nomenclature. When a_5 is proline, this is the only possible structure which maintains a normal β-bend structure for residues $a_1 - a_4$. However, in the group II repeats it is possible for a_5 (Ala or Gly) to assume the x position, while a_1 and a_2 become 1 and 2, thus transferring the bulge from the C-terminal to the N-terminal side of the bend. The bend will still have its hydrophobic face tilted towards the hydrophobic face of the sheet.

The tilted bends, which exclude water from the hydrophobic core, perform a homologous function to the cross-over conditions between the β-sheets of β-sandwich domains with a Greek key topology (Richardson, 1981). We looked for evidence in globular proteins of known structure that five-residue bends might also serve this function. A survey of 62 proteins found 24 five-residue bends between anti-parallel strands, two with prolines in position 5, but none possessed our proposed configuration (J.M. Thornton, unpublished work). At first sight, this absence is unexpected but it may be correlated with the unusual shape of the fibre protein.

A structure, very similar in its general features to our model, was previously proposed by Earnshaw *et al.* (1979) for the gene 37 protein, which forms the distal half of the tail fibre of bacteriophage T_4 but the absence of clear repeating sequences prevented precise conclusions.

Folding and assembly of the dimeric fibre

Although the proposed structure is plausible, some difficulties arise when the folding of the newly synthesised polypeptide chain is considered. Most proteins fold so as to minimise the exposure of hydrophobic groups to water both during the folding process and in the final product (Ptitsyn and Finkelstein, 1980). The only extensive exposure of such groups is likely to be at subunit interfaces. In the monomeric fibre protein this interface would run the whole length of the shaft. It is unlikely that such a long narrow sheet would be stable. A more likely possibility is that the polypeptide chain of the shaft adopts a more or less random coil structure until it meets its complementary subunit and that the folding process is nucleated by the much longer hydrophobic strands of the knob. Dimerisation of the knob would bring the segments of the shaft into register, whereupon they would fold, burying their hydrophobicity between them (or in other words 'leave them alone and they will come home, twisting their tails behind them'). This would also solve the problem of correct dimerisation for a repetitious surface.

This analysis of the folding process suggested an alternative structure. The twin tails could fold in parallel to form an intertwined double helical structure of three residue β-strands linked by bends. This flattened 'helix' would have a hydrophobic core, the hydrogen bonding between the strands would be parallel rather than anti-parallel and the β-bends would be almost perpendicular to the axis of the helix. The topology of the strand connections (Figure 7) is similar to that of a 'Swiss roll β-barrel', such as is found in tomato bushy stunt virus, except that the latter is derived from an antiparallel hairpin rather than from parallel chains (Figure 7c). The double helical structure (Figure 7b) can be formed by winding the polypeptide chains in either a left or right handed sense. The right handed helix is more probable because of the almost exclusive preference for right hand crossover connections observed in globular proteins (Richardson, 1981). This would also be consistent with the usual right handed twist of each β-strand (Chothia, 1973).

When we built models of this structure the bulk of the hydrophobic residues made it difficult to join the two sheets by the four- or five-residue bends. However, it was just possible to build the structure with the CPK models so that one cannot exclude it on steric grounds, although it did appear improbably highly strained. Moreover, it was necessary to connect the strands, as shown in Figure 7b, with a right-handed cross-over between the opposing sheets. In contrast, a left-handed cross-over is always found in globular proteins (e.g., Figure 7c) because this maintains maximal side chain contact over the whole of a twisted β-sheet (Chothia and Janin, 1981).

The enforced cross-over of ~20° between the strands on the opposite sides of the double helical structure provides possible ground for a distinction from the anti-parallel, side-by-side dimer in which the two sets of strands are likely to be almost parallel. The 20° angle between the strands would be

expected to lead to a splitting, or at least a spreading, of the meridional arcs of the electron diffraction pattern over ~30°. This is significantly greater than the observed spread of 18°, which therefore favours the anti-parallel structure. A strong argument against the double helical structure is that in it the unusual five-residue bends have no obvious function, whereas they make an important contribution to the stability of our proposed structure.

The dimeric structure of the fibre, which follows both from our proposals and from the results of Devaux *et al.* (see Acknowledgements), adds a further symmetry problem to those which already have to be considered in the assembly of the capsid of the adenovirus. The trimeric penton base (Devaux *et al.*, 1982) has to associate on the one hand with the fibre and on the other with the five hexons which surround it. The polyoma virus capsid provides another recent example of a departure from the principles of quasi-equivalence (Rayment *et al.*, 1982).

Materials and methods

Purification and crystallisation of adenovirus fibre

KB or HeLa cells growing in suspension were infected with adenovirus type 5 as described previously (Winters and Russell, 1971). Infected cells from ~8 litres of suspension culture were disrupted by treatment with a fluorocarbon (Arklone L, ICI Ltd.) and the extracts submitted to velocity gradient centrifugation in caesium chloride, producing an opalescent virus band well separated from an upper layer of excess virus structural components (Winters and Russell, 1971; Russell *et al.*, 1967). The latter region of the gradient (consisting mainly of hexon, penton and fibre capsomeres) was dialysed against 50 mM HEPES NaOH buffer pH 6.8 and chromatographed on QAE Sephadex A-25 (Pharmacia). After washing with the above buffer, the first fractions eluted by 0.25 M NaCl in buffer gave fractions consisting essentially of fibre protein, as determined by gel electrophoresis. The appropriate fractions were pooled and concentrated by vacuum dialysis against 50 mM phosphate buffer pH 6.8 to a volume of ~1–2 ml. The concentrate was then dialysed at 4°C against a large volume of 10 mM phosphate buffer pH 6.0 (Mautner and Pereira, 1971). Crystals of adenovirus fibre usually formed after 1–2 days and could be centrifuged at 500 *g* for 5 min and washed in pH 6.0 buffer and then resuspended in the same buffer.

Samples were prepared for gel electrophoresis by boiling for 2 min in SDS, urea and mercaptoethanol before analysis on 10% or 15% polyacrylamide slab gels as described previously (Russell and Blair, 1977).

Protein concentration was determined from the u.v. absorption spectrum using a value of $E_{1\,cm}^{1\%}$ of 8.5, determined from duplicate amino acid analyses on two different preparations. The analyses were performed on 18 h hydrolysates of the fibre protein (6 M HCl, 0.1% phenol 110°C under N_2) using a Beckman 121M amino acid analyser. Tryptophan/tyrosine ratios were determined from the u.v. spectrum (Melamed and Green, 1963) and used to calculate the tryptophan content.

C.d. spectra were recorded digitally from 270 nm to 196 nm with a Jasco J41 C spectropolarimeter equipped with a J-DPY data processor (sensitivity 0.5 millidegrees/cm, time constant 4 or 16). Protein samples 0.12 mg/ml in 50 mM sodium borate, pH 9.2) were transferred to 0.1 cm cuvettes and the spectrum, averaged over six scans, was recorded at 22°C. The u.v. spectrum was determined on the same sample with a Cary 219 spectrophotometer. The c.d. spectra are expressed as molar $\Delta\epsilon$ based on a mean residue weight of 110. The analysis of the secondary structure of the protein was performed by the method of Provencher and Glöckner (1981) in which the experimental curve is matched to a linear combination of the c.d. spectra of 16 proteins of known structure.

Electron micrographs of single molecules and of crystals were obtained from samples adsorbed on carbon films and stained with 1% sodium silicotungstate. For electron diffraction measurements, crystals suspended in phosphate buffer (0.01 M pH 6.2) were air dried on carbon films without stain and with or without addition of 1% glucose. Diffraction patterns were recorded in a JEM 100C microscope at 100 keV, using the objective as a third illuminating lens. Low angle reflections were stable but the 4.7 Å cross-β reflection was very short lived and its detection required extreme low-dose search procedures (Wrigley *et al.*, 1983). The diffraction patterns were calibrated using powder diffraction from gold, evaporated on to a separate support. A low magnification image of the diffracting crystal was recorded in a second exposure on the same negative, to give the angular relation between the diffraction pattern and the fibre molecules.

The programme of Roberts and Geisow (1979) based on the method of Chou and Fasman (1978) was used to locate regions of potential α-helix, β-sheet and β-bend. The programme also provided a plot of mean hydrophobicity of four residue segments which proved particularly useful.

Acknowledgements

We thank B. Precious, R.K. Chillingworth, E. Brown, E.J. Toms and S. Lathwell for excellent technical assistance and Dr. V. Mautner, Dr. J.M. Thornton and Professor T. Blundell, who provided stimulating discussions. We also thank Drs. C. Devaux, C. Berthet-Colominas, P.A. Timmins and B. Jacrot (Institut Laue-Langevin, Grenoble) for allowing us to quote from their 1983 results in advance of publication.

References

Boudin,M. and Boulanger,P. (1982) *Virology,* **116**, 589-609.
Boulanger,P. and Loucheux,M.H. (1972) *Biochem. Biophys. Res. Commun.,* **47**, 194-201.
Chothia,C. (1973) *J. Mol. Biol.,* **75**, 295-302.
Chothia,C. and Janin,J. (1981) *Proc. Natl. Acad. Sci. USA,* **78**, 4146-4150.
Chou,P.Y. and Fasman,G.D. (1978) *Adv. Enzymol.,* **47**, 45-140.
Dayhoff,M. (1978) in *Atlas of Protein Sequence and Structure.* Vol. **5**, supplement 3, published by National Biomedical Research Foundation, Washington, DC, pp. 345-352.
Devaux,C., Zulauf,M., Boulanger,P. and Jacrot,B. (1982) *J. Mol. Biol.,* **156**, 927-937.
Dorsett,P.H. and Ginsberg,H.S. (1975) *J. Virol.,* **15**, 208-216.
Earnshaw,W.C., Goldberg,E.B. and Crowther,R.A. (1979) *J. Mol. Biol.,* **132**, 101-131.
Fitch,W.M. (1977) *Genetics,* **86**, 623-644.
Geddes,A.J., Parker,K.D., Atkins,E.D.T. and Beighton,E. (1968) *J. Mol. Biol.,* **32**, 343-358.
Gelderblom,H. and Maichle-Laupe,I. (1982) *Arch. Virol.,* **72**, 289-296.
Ginsberg,H. (1979) in Fraenkel-Conrat,H. and Wagner,C. (eds.), *Comparative Virology,* Vol. **14**, Plenum Press, NY, pp. 409-427.
Green,M., Mackey,K., Wold,W.S.M. and Rigden,P. (1979) *Virology,* **93**, 481-492.
Hérissé,J. and Galibert,F. (1981) *Nucleic Acids Res.,* **9**, 1229-1240.
Hérissé,J., Rigolet,M., de Dinechin,S.D. and Galibert,M. (1981) *Nucleic Acids Res.,* **9**, 4023-4042.
Laver,W.G., Younghusband,H.B. and Wrigley,N.G. (1971) *Virology,* **45**, 598-614.
McLachlan,A.D. (1971) *J. Mol. Biol.,* **61**, 409-424.
McLachlan,A.D. (1972) *J. Mol. Biol.,* **64**, 417-437.
Mautner,V. and Pereira,H.G. (1971) *Nature,* **230**, 456-457.
Melamed,M.D. and Green,N.M. (1963) *Biochem. J.,* **89**, 591-599.
Norrby,E. (1969) *J. Gen. Virol.,* **5**, 221-236.
Provencher,S.W. and Glöckner,J. (1981) *Biochem. J.,* **20**, 33-37.
Ptitsyn,O.B. and Finkelstein,A.V. (1980) *Q. Rev. Biophys.,* **13**, 339-386.
Rayment,I., Baker,T.S., Caspar,D.L.D. and Murakami,W.T. (1982) *Nature,* **295**, 110-115.
Richardson,J.S. (1981) *Adv. Protein Chem.,* **34**, 167-339.
Richardson,J.S., Getzoff,E.D. and Richardson,D.C. (1978) *Proc. Natl. Acad. Sci. USA,* **75**, 2574-2578.
Roberts,R.D.B. and Geisow,M. (1980) *Int. J. Macromol.,* **2**, 387-389.
Russell,W.C. and Blair,G.E. (1977) *J. Gen. Virol.,* **34**, 19-35.
Russell,W.C., Hayashi,K., Sanderson,P.J. and Pereira,H.G. (1967) *J. Gen. Virol.,* **1**, 495-507.
Sundquist,B., Pettersson,U., Thelander,L. and Phillipson,L. (1973) *Virology,* **51**, 252-256.
Valentine,R.C. and Pereira,H.G. (1965) *J. Mol. Biol.,* **13**, 13-22.
Winters,W. and Russell,W.C. (1971) *J. Gen. Virol.,* **10**, 181-194.
Wadell,G. (1979) *Intervirology,* **11**, 47-57.
Woody,R.W. (1974) in Blout,E.R., Bovey,F.A., Goodman,M. and Lotan,N. (eds.), *Peptides, Polypeptides and Proteins,* John Wiley, NY.
Wrigley,N.G., Brown,E. and Chillingworth,R.K. (1983) *J. Microsc.,* **130**, 225-232.